Paleobiology

cambridge.org/pab

The FossilSimShiny app for simulating under the fossilized birth–death process and plotting the results

Titouan Chambe[1], Rachel C. M. Warnock[2] and Joëlle Barido-Sottani[3]

[1]Université Paris-Saclay, 91190, Gif-sur-Yvette, France
[2]GeoZentrum Nordbayern, Department of Geography and Geosciences, Friedrich-Alexander Universität Erlangen-Nürnberg, Erlangen, Germany
[3]Institut de Biologie de l'ENS (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

Methodological Advances

Cite this article: Chambe T, Warnock RCM, Barido-Sottani J (2024). The FossilSimShiny app for simulating under the fossilized birth–death process and plotting the results. *Paleobiology* **50**, 401–407. https://doi.org/10.1017/pab.2024.34

Received: 13 February 2024
Revised: 17 July 2024
Accepted: 30 July 2024

Corresponding author:
Joëlle Barido-Sottani;
Email: joelle.barido-sottani@bio.ens.psl.eu

Non-technical Summary

Integrating information from the fossil record allows us to obtain accurate age estimates for important events in the evolutionary history of living species and to study the dynamics of speciation and extinction in more detail. The fossilized birth–death (FBD) process is a widely used model for the diversification of both extinct and extant species. Understanding the behavior of this model and the influence of its different parameters is thus very important for researchers interested in reconstructing the evolutionary history of species through time. Here, we present a web application built around the FBD process that allows users to simulate diversification and fossil sampling and visualize the results. The app can be used as a demonstration and teaching tool, allowing users to experiment with the different components of the model and observe their effect on the results of the simulation. It also helps researchers in designing simulation studies and testing their simulation choices.

Abstract

Phylogenetic inferences using combined datasets of both extant and extinct species have grown increasingly popular, in part thanks to the development of the fossilized birth–death (FBD) process. The FBD process provides a powerful model for the evolution of past and present lineages and can be used for both inference and simulation. Simulations in particular are very helpful for new users to gain better understanding of the model and its different components. In this work, we present FossilSimShiny, a visual application for simulating phylogenies, fossil samples, and fossil taxonomies under the FBD process. The app integrates a wide range of simulation models and presents the simulation results in clear, customizable figures. As a teaching tool, FossilSimShiny allows lecturers to create illustration plots and students to directly experiment with the model. For research applications, the app can help researchers save time and effort by testing and calibrating simulation setups before running them on a large scale.

Introduction

Phylogenetic trees allow us to represent the evolutionary relationships between organisms and species and to obtain information about the underlying diversification processes. Molecular sequences are commonly used to estimate phylogenies; however, these sequences are generally only available for extant species. To obtain information about past dynamics and to better estimate the divergence times, it is necessary to use temporal information, often provided by the fossil record. One model developed for this purpose if the fossilized birth–death (FBD) process, which integrates both fossil and extant specimens into a combined phylogeny. This model provides a statistically coherent tree prior for Bayesian phylogenetic inference (Stadler 2010; Heath et al. 2014) and has been widely used in empirical studies to obtain complete representations of the evolutionary process through time (e.g., Thomas et al. 2020; Pohle et al. 2022). The model has since been extended in different ways to allow for variation rates through time and across taxa (Gavryushkina et al. 2014; Kühnert et al. 2016). As a birth–death model, the FBD process can also be used as a forward simulation model, for instance, in the R package FossilSim (Barido-Sottani et al. 2019b). This package can generate complete or reconstructed phylogenies under the FBD process and integrates different models for representing variations in the fossil sampling process. FossilSim can also simulate the taxonomy of fossil species, which describes how fossils are sorted in the record on the basis of morphology and is a crucial component of empirical datasets.

Simulation tools such as FossilSim have many applications. Large simulated datasets are useful for validating FBD inferences (e.g., Barido-Sottani et al. 2019a), but individual

© The Author(s), 2024. Published by Cambridge University Press on behalf of Paleontological Society. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

PALEOBIOLOGY A PUBLICATION OF THE Paleontological SOCIETY

CAMBRIDGE UNIVERSITY PRESS

simulations also provide an ideal opportunity for users to explore the behavior of the model under different conditions and parameters, and thus understand it better. The rapid expansion of the FBD family of models and the myriad ways in which the model can be applied can make it challenging for new empirical users to choose the set up that is most appropriate for their data. Here we present the FossilSimShiny app, a web app which provides an intuitive and accessible interface to the simulation and plotting functions of the FossilSim package. Through the app, users can easily visualize the outcome of simulations under the FBD process, and the influence of the different options on the results. Teachers can use FossilSimShiny to generate custom illustrations for presentations or courses, or directly make it available for students to experiment with the different components of the model. Because FossilSimShiny is itself an R package, it is easy to install and run, either as a standalone tool on a user's machine, or on a server for broader access.

FossilSimShiny

Getting Started

FossilSimShiny is available as an R package on the CRAN repository (https://cran.r-project.org/package=FossilSimShiny). It requires a working installation of R and can be downloaded and installed using the command

```
install.packages("FossilSimShiny")
```

in any R console. Once installed, the command

```
FossilSimShiny::launchFossilSimShiny()
```

will start the app locally in the default browser. The app can also be installed on a server in order to make it accessible to a wider audience, for instance, for teaching purposes. Detailed instructions on server installation can be found in the package documentation.

The landing page of the app, shown in Figure 1, contains three simulation submenus covering the tree, taxonomy, and fossils. The fourth submenu contains options to change the appearance of the plots. Finally, the app allows the simulated data and the generated plots to be downloaded for future reference. Hovering over each parameter or option will show additional information in a tooltip at the bottom of the screen.

Simulation

The first step in using the app consists of simulating data, using the simulation submenus:

1. Tree: Phylogenies are simulated using a simple birth–death process conditioned on the number of taxa at present. The user needs to specify the birth and death rates used, as well as the number of extant taxa. The app also allows the user to

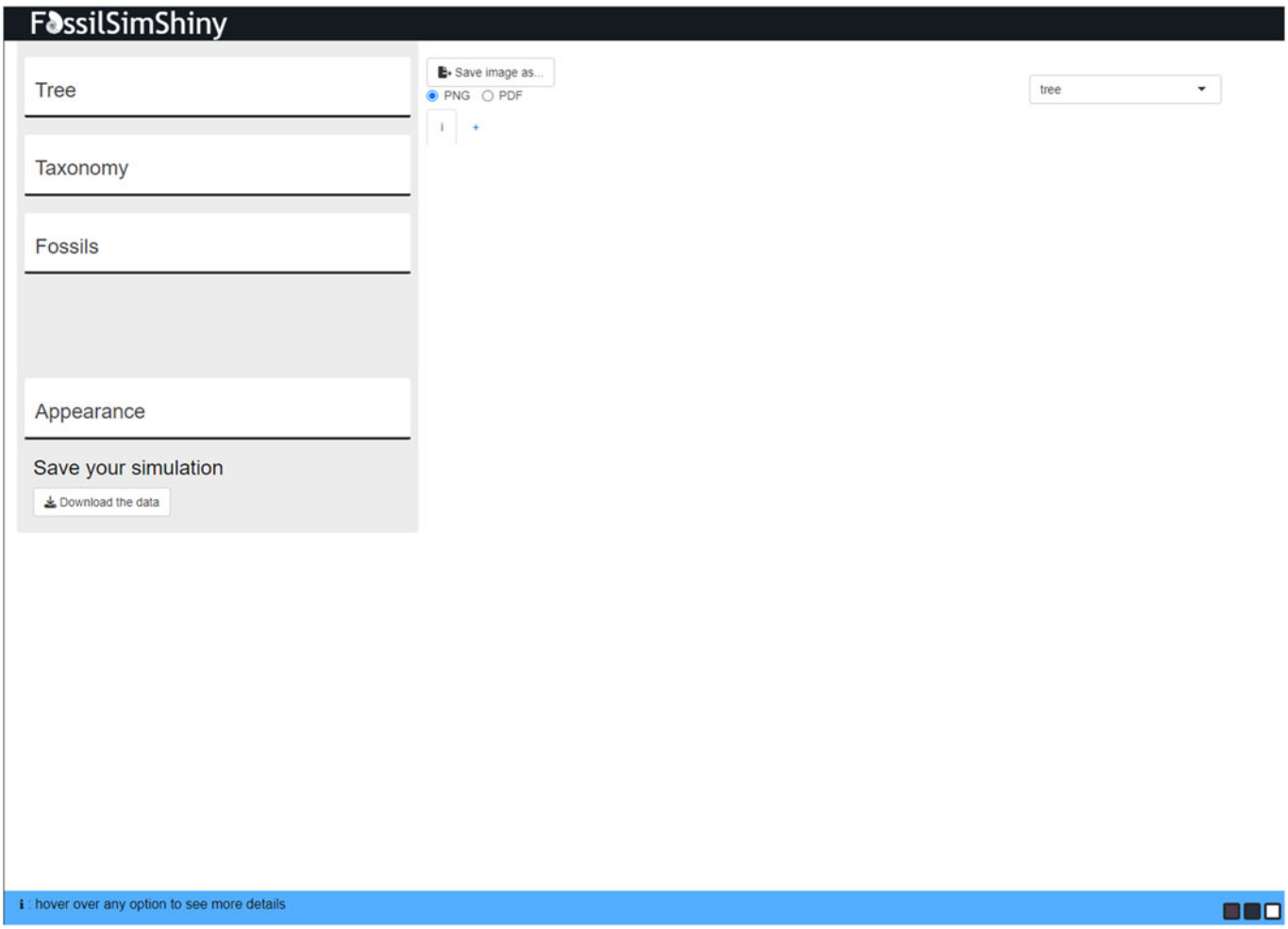


Figure 1. Main menu of the FossilSimShiny app.

https://doi.org/10.1017/pab.2024.34 Published online by Cambridge University Press

provide a chosen rooted tree in Newick format instead. Once simulated or imported, the full tree is automatically plotted by the app, as shown in Figure 2.

2. Taxonomy: The taxonomy is simulated based on the phylogeny, using the mixed-speciation model presented in Stadler et al. (2018). This model represents how fossil species are classified in the fossil record and thus decouples the origin of "morphospecies" from branching events in the tree. It accounts for bifurcating, budding, and anagenetic speciation events. Once simulated, the taxonomy is automatically plotted by the app, as shown in Figure 3.
3. Fossils: Fossil sampling is simulated based on the phylogeny. Several fossil sampling models are available, including uniform sampling across the tree, time-dependent sampling, environment-dependent sampling, and lineage-dependent sampling. Time-dependent sampling is represented as a piecewise-constant process, also known as a "skyline" or "episodic" model, wherein the rates follow a lognormal distribution specified by the user. Environment-dependent sampling follows the model presented in Holland (1995), wherein fossil sampling rates depend on an environmental proxy combined with lineage-specific environmental preferences. Finally, lineage-dependent sampling simulates edge-specific sampling rates drawn from a lognormal distribution specified by the user. If a taxonomy has been simulated for the phylogeny, it will be used to simulate rates based on the species rather than the edges. Once simulated, the full tree including the fossil samples will be automatically plotted by the app, as shown in Figure 4.

All simulation functions will also print the amount of time taken for the simulation on top of the plot. In addition, the taxonomy simulation will print the number of bifurcating, budding, and anagenetic events simulated, and the fossil simulation will print the number of simulated fossil samples.

Plotting

The app contains three main plotting options: tree alone, taxonomy, and tree with fossils. When simulating, the option is automatically switched to the one corresponding to the simulation. The different options can also be selected manually using a dropdown menu.

The Appearance submenu provides additional plotting options to precisely control the appearance of the final plot. For example, the user can choose to plot the reconstructed tree instead of the full tree, showing only the lineages that lead to fossil or extant samples. Numbered tip labels can also be added to the plot. Some options are only available if fossil samples have been simulated first, such as showing the fossil species as ranges instead of individual specimens or showing the fossils alone without the underlying phylogeny. Finally, some options are specific to

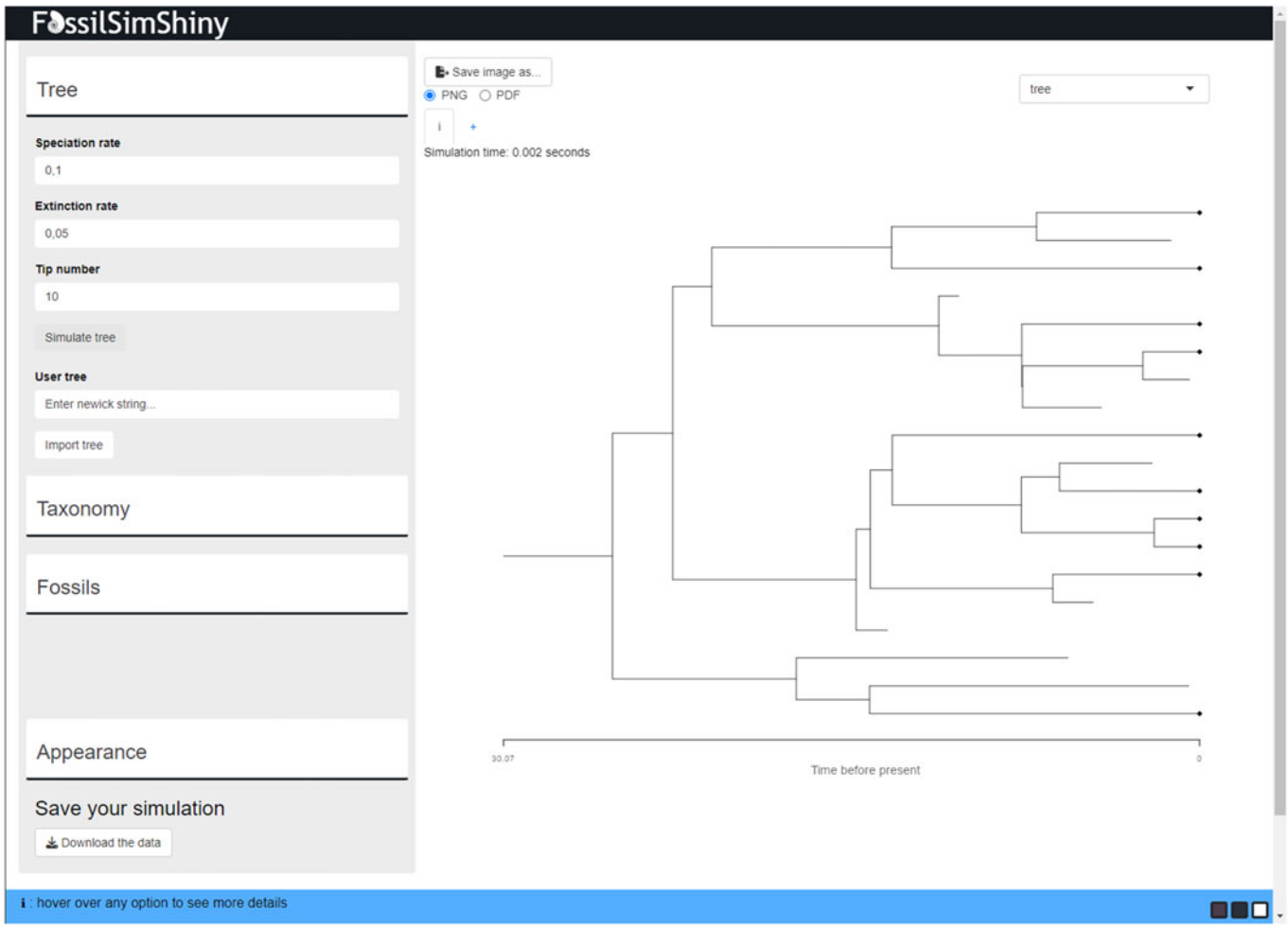


Figure 2. Tree simulation in the FossilSimShiny app. By default, the full tree is shown, including extinct lineages.

https://doi.org/10.1017/pab.2024.34 Published online by Cambridge University Press

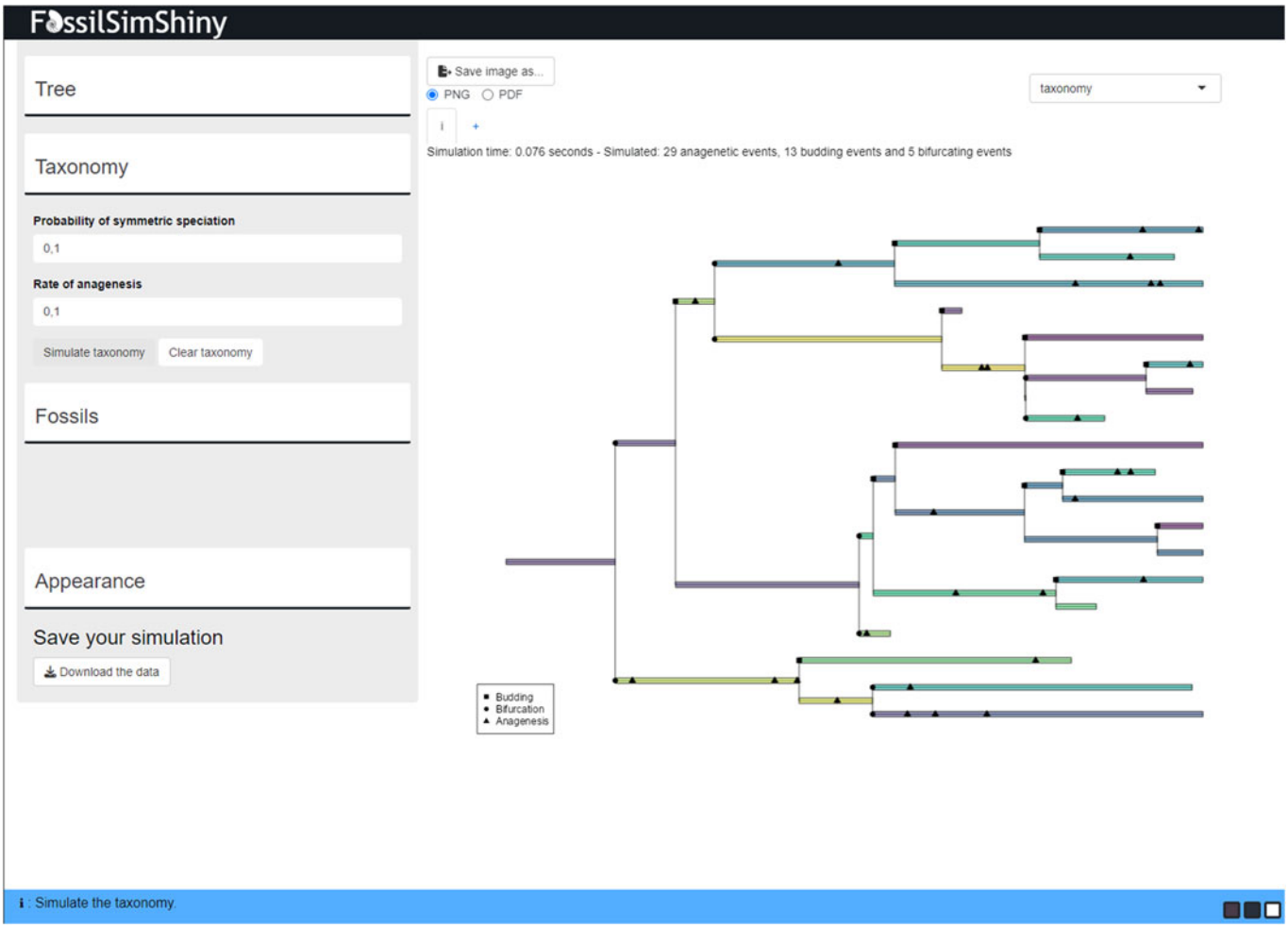


Figure 3. Taxonomy simulation in the FossilSimShiny app, showing the simulated species as color ranges. Each speciation event is labeled with the mode of speciation: budding, bifurcating, or anagenetic.

certain simulation models, for instance, the time intervals used for time-dependent sampling or the environmental variables used for environment-dependent sampling. A summary of all currently available options is shown in Table 1.

Multiple Plots

The app includes a tab system that allows users to run and plot several simulations in parallel. Each tab contains its own tree, taxonomy, and fossil samples, which will be saved when switching to another tab or opening a new one. This allows users to easily compare the results obtained from different setups, as shown in Figure 5. The app can currently support up to five simultaneous simulations.

Exporting the Simulations

Data from the app can be exported in two separate ways. First, the plot can be downloaded in PNG or PDF format in order to be included in a paper or presentation. This function will save the plot in the currently selected tab with the currently selected appearance options, exactly as it appears in the app. The second possibility is to directly save the simulated data as an RData file. The downloaded data will contain the simulated phylogeny in the phylo format used by most phylogenetics packages, and the simulated taxonomy and fossil specimens in the formats used by the package FossilSim. The resulting file can be loaded easily via R to perform further simulations or to plot with additional options that are not available through the app.

Technical Implementation

FossilSimShiny is built using Shiny (https://shiny.rstudio.com), an R package that allows web apps to be developed using R code. It also uses Javascript code to perform some functions more quickly, such as showing help on the different configuration options. As a backend, FossilSimShiny relies on the R packages TreeSim (Stadler 2011) for simulating phylogenies and FossilSim (Barido-Sottani et al. 2019b) for simulating taxonomies and fossils, as well as plotting all output.

Research Applications

The underlying package FossilSim has been used in many simulation studies (e.g., Barido-Sottani et al. 2019a; Černý et al. 2021). One of the difficulties that can be encountered in such simulation studies is to calibrate the parameters, such as the birth, death, and fossilization rates, to obtain datasets with the desired

https://doi.org/10.1017/pab.2024.34 Published online by Cambridge University Press

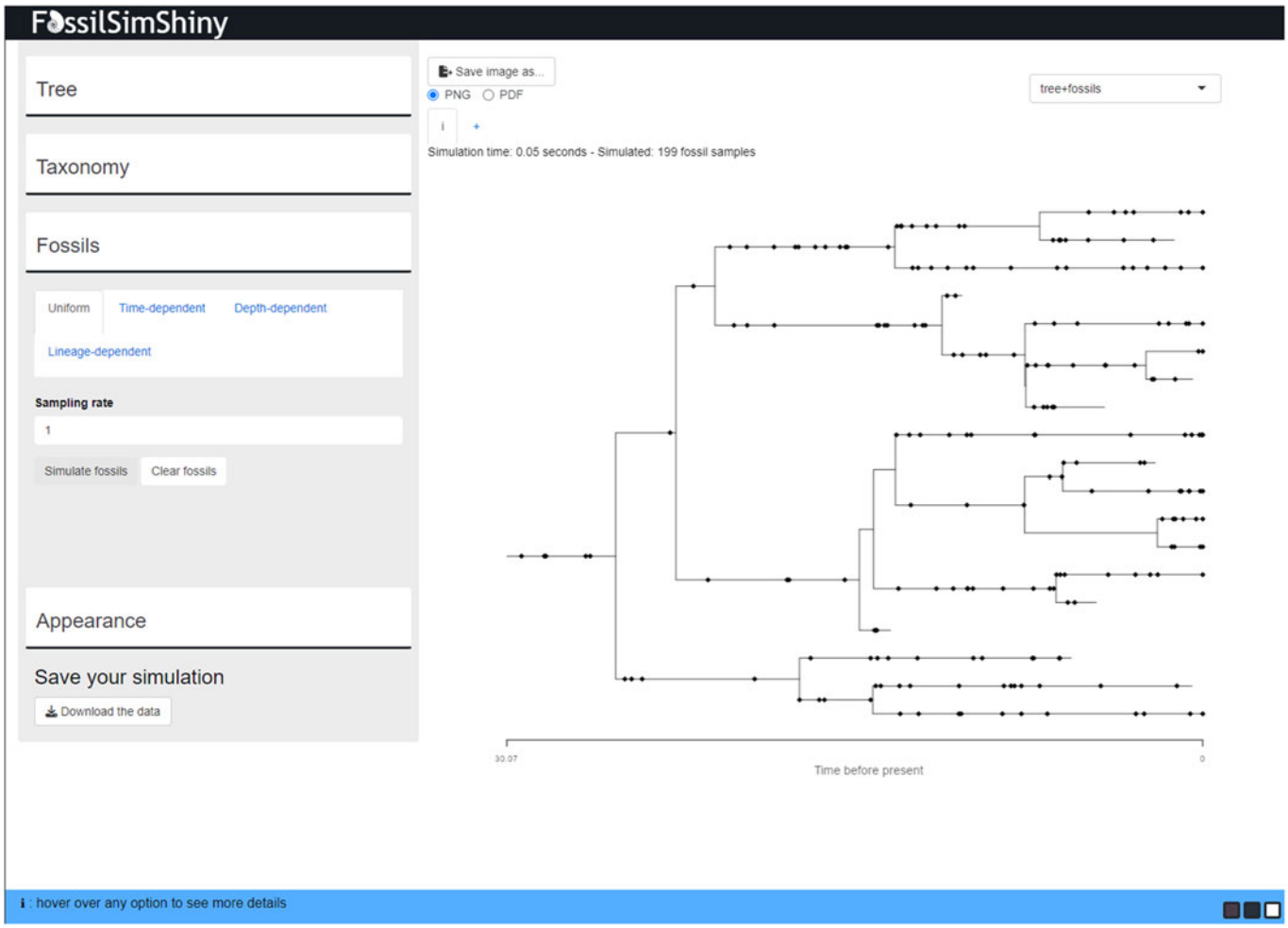


Figure 4. Fossil sampling simulation in the FossilSimShiny app using the uniform sampling model. Each dot represents a fossil sample.

characteristics. For instance, simulation studies will frequently target a specific range for the root ages or the total number of tips (extant and fossil) for their simulated phylogenies. This ensures that the simulated replicates are large enough to be representative of an empirical dataset, but small enough to limit the computational cost of the study. It also allows the replicates to be more directly comparable, as certain output metrics can be influenced by tree size. For instance, some measures of topological distance between inferred and true trees rely on counting splits, but the number of possible splits for a given dataset is dependent on the number of samples. However, it is not always straightforward to choose parameter values to obtain the desired result, in particular for more complex models, where the parameters can interact in unexpected ways. The final number of recorded fossil samples, for instance, depends on the interaction between the birth and death rates, the age of the phylogeny, and the fossil sampling model and parameters. In general, higher birth and fossilization rates, lower extinction rates, and a higher tree age will lead to

Table 1. Summary of FossilSimShiny plotting options and their requirements.

Option	Description	Requirements
Show tree	Show/hide the phylogeny	None
Show fossil taxonomy	Show/hide the species assignment of fossil samples	Fossil samples and taxonomy
Show all occurrences	Show/hide all individual fossil samples	Fossil samples
Show ranges	Show/hide ranges between first and last occurrences of a given lineage or species	Fossil samples
Show strata	Show/hide the time intervals used for fossil simulation	None
Show tip labels	Show/hide tip names for all tips in the phylogeny	None
Show reconstructed tree	Show the reconstructed tree without unsampled lineages or the full tree	None
Show depth used in sampling	Show depth values used for depth-dependent fossil sampling	Fossil samples under a depth-dependent model

https://doi.org/10.1017/pab.2024.34 Published online by Cambridge University Press

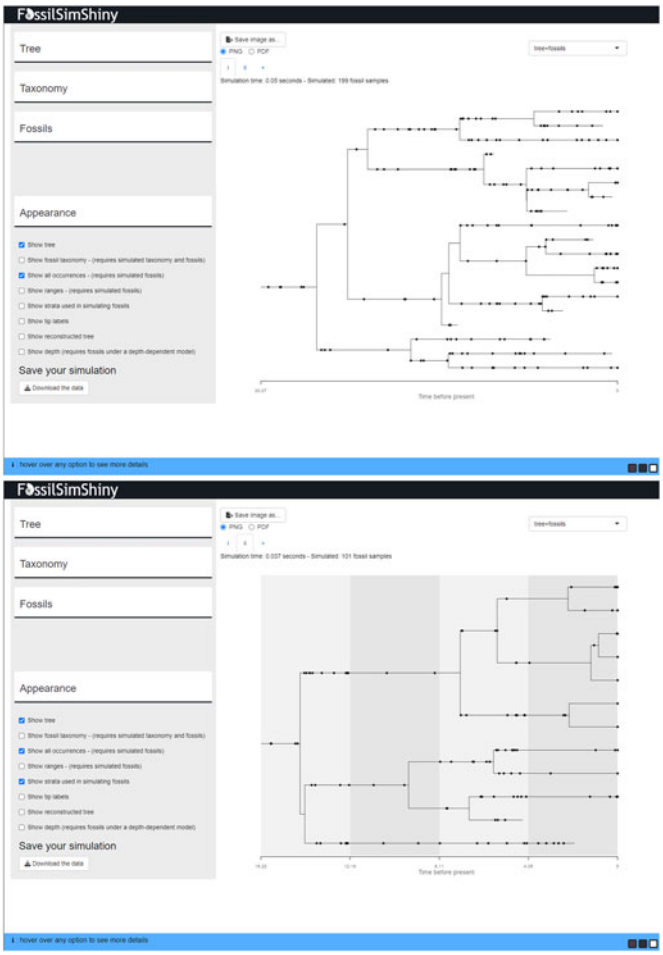


Figure 5. Example of two concurrent simulations plotted in two different tabs of FossilSimShiny.

greater numbers of recorded fossils, but these general trends can be difficult to translate directly into usable parameter values. FossilSimShiny helps users test and pick appropriate parameter values based on the desired features of the simulated dataset. As the complexity of models grows, the potential for unexpected interactions between the different components and thus of undesirable simulation outcomes also expands. For instance, a simulation setup intended to generate within-lineage heterogeneity can, depending on the chosen setup and parameter values, lead to datasets in which most replicates are homogeneous, completely defeating the purpose of the simulations. One way this can happen is that if there is too much discrepancy between the fossilization rates of different lineages, lineages with low rates may not be represented by any samples in the final dataset. Alternatively, if the process of transitioning between heterogeneous categories is too slow, lineages of the tree may all stay in the initial category initiated at the root. Such issues can be difficult to anticipate and only become apparent when observing the simulation outcomes. By doing a test run in FossilSimShiny, researchers can identify problematic behaviors in advance and can integrate the appropriate corrections or validation steps into their simulation pipeline. Overall, FossilSimShiny allows researchers to quickly and efficiently test a simulation setup on a smaller scale, before spending large amounts of computation time on simulating a full-size dataset.

Conclusions

The FossilSimShiny app provides an intuitive and easily accessible interface to perform simulations under the FBD process. As shown in our example tutorial, it allows students and new users of phylogenetic models to visualize the impact of different parameters and conditions on the output and thus to gain a better understanding of the model behavior. In addition, FossilSimShiny can be used easily to produce example plots for scientific presentations or teaching purposes, while accurately representing the dynamics of the FBD process. Finally, FossilSimShiny can be used by researchers to calibrate simulation parameters and check their setups for unexpected outcomes before running the full pipeline, saving both researcher time and computation time.

Future work on the app will integrate more of the available options in FossilSim, including additional models for fossil sampling and further options for customizing different plots. We will also expand the import options to allow users to import and plot their own simulated data. Other features will be implemented based on user feedback. Indeed, we encourage users of FossilSimShiny to send us bug reports and feature requests by filing an issue on our GitHub repository (https://github.com/fossilsim/shiny/issues).

Acknowledgments. JBS was supported by funds from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 101022928. We thank the Morlon team, T. J. Smith, and W. Gearty for providing feedback on FossilSimShiny and on this article.

Competing Interest. The authors declare no competing interests.

Data Availability Statement. The full source code is freely available on GitHub (https://github.com/fossilsim/shiny). The app is also available as a package on CRAN (https://cran.r-project.org/package=FossilSimShiny). The app can be run locally using the instructions in the package or can be installed on a server using the instructions in the vignette "Hosting FossilSimShiny on a Web Server."

The latest release of FossilSimShiny 1.1.2 is currently hosted on the Shiny server (https://fossilsim.shinyapps.io/shinyapp) and is freely accessible to users. We provide an example tutorial to demonstrate how the app can be used for teaching (https://phylogenetics-fau.netlify.app/fossilsimshiny).

Literature Cited

Barido-Sottani, J., G. Aguirre-Fernández, M. J. Hopkins, T. Stadler, and R. Warnock. 2019a. Ignoring stratigraphic age uncertainty leads to erroneous estimates of species divergence times under the fossilized birth-death process. *Proceedings of the Royal Society B* **286**:20190685.

Barido-Sottani, J., W. Pett, J. E. O'Reilly, and R. C. M. Warnock. 2019b. FossilSim: an R package for simulating fossil occurrence data under mechanistic models of preservation and recovery. *Methods in Ecology and Evolution* **10**:835–840.

Černý, D., D. Madzia, and G. J. Slater. 2021. Empirical and methodological challenges to the model-based inference of diversification rates in extinct clades. *Systematic Biology* **71**:153–171.

Gavryushkina, A., D. Welch, T. Stadler, and A. J. Drummond. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Computational Biology* **10**:e1003919.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences USA* **111**:E2957–E2966.

Holland, S. M. 1995. The stratigraphic distribution of fossils. *Paleobiology* **21**:92–109.

https://doi.org/10.1017/pab.2024.34 Published online by Cambridge University Press

Kühnert, D., T. Stadler, T. G. Vaughan, and A. J. Drummond. 2016. Phylodynamics with migration: a computational framework to quantify population structure from genomic data. *Molecular Biology and Evolution* **33**:2102–2116.

Pohle, A., B. Kröger, R. Warnock, A. H King, D. H. Evans, M. Aubrechtová, M. Cichowolski, X. Fang, and C. Klug. 2022. Early cephalopod evolution clarified through Bayesian phylogenetic inference. *BMC Biology* **20**:88.

Stadler, T. 2010. Sampling-through-time in birth–death trees. *Journal of Theoretical Biology* **267**:396–404.

Stadler, T. 2011. Simulating trees with a fixed number of extant species. *Systematic Biology* **60**:676–684.

Stadler, T., A. Gavryushkina, R. C. M. Warnock, A. J. Drummond, and T. A. Heath. 2018. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. *Journal of Theoretical Biology* **447**:41–55.

Thomas, D. B., A. J. D. Tennyson, R. P. Scofield, T. A. Heath, W. Pett, and D. T. Ksepka. 2020. Ancient crested penguin constrains timing of recruitment into seabird hotspot. *Proceedings of the Royal Society B* **287**:20201497.

https://doi.org/10.1017/pab.2024.34 Published online by Cambridge University Press